Fax-On-Demand
Telephone: (202)401-0527
Item: 6023

LIST OF ATTACHMENTS

ATTACHMENT 1

ATTACHMENT 1: Glossary

Beta Distribution is a flexible, bounded PDF described by two shape parameters. It is commonly used when a range of the random variable is known. (p. A3-14)

Boxplot is a graphical representation showing the center and spread of a distribution, along with a display of outliers. (p. A3-10)

Central Limit Theorem says that for a relatively large sample size, the random variable x (the mean of the samples) is normally distributed, regardless of the population's distribution. (p. A3-14)

Coefficient of Variation (also Coefficient of Variance or Coefficient of Variability)* is an estimate of relative standard deviation. Equals the standard deviation divided by the mean. Results can be represented in percentages for comparison purposes. (p. A3-7)

Confidence Interval is the range within which one has a given level of confidence that the range includes the true value of the unknown parameter (e.g. a 95% confidence interval for a parameter means that 95% of the time the true value of that parameter will be within the interval).

Continuous Probability Distribution* is a probability distribution that describes a set of uninterrupted values over a range. In contrast to the Discrete distribution, the Continuous distribution assumes there are an infinite number of possible values.

Correlation, Correlation Analysis is an investigation of the measure of statistical association among random variables based on samples. Widely used measures include the linear correlation coefficient (also called the product-moment correlation coefficient or Pearson correlation coefficient), and such non-parametric measures as Spearman rank-order correlation coefficient, and Kendall's tau. When the data are nonlinear, non-parametric correlation is generally considered to be more robust than linear correlation.

Correlation Coefficient* is a number between -1 and 1 that specifies mathematically the degree of positive or negative correlation between assumption cells. A correlation of 1 indicates a perfect positive correlation, minus 1 indicates a perfect negative correlation, and 0 indicates there is no correlation.

Cumulative Distribution Function (CDF) is alternatively referred to in the literature as the distribution function, cumulative frequency function, or the cumulative probability function. The cumulative distribution function, F(x), expresses the probability the random variable X assumes a value less than or equal to some value x, $F(x)=\text{Prob}(X \leq x)$. For continuous random variables, the cumulative distribution function is obtained from the probability density function by integration. In the case of discrete random variables, it is obtained by summation.

Cumulative Frequency Distribution is a chart that shows the number or proportion (or percentage) of values less than or equal to a given amount.

Deterministic Model, as opposed to a stochastic model, is one which contains no random elements.

Discrete Probability Distribution* is a probability distribution that describes distinct values, usually integers, with no intermediate values. In contrast, the continuous distribution assumes there are an infinite number of possible values.

Distribution is the pattern of variation of a random variable.

Frequency (also Frequency Count)* is the number of times a value recurs in a group interval.

Frequency Distribution* is a chart that graphically summarizes a list of values by subdividing them into groups and displaying their frequency counts.

Goodness-of-Fit is a set of mathematical tests performed to find the best fit between a standard probability distribution and a data set.

Goodness-of-Fit Test is a formal way to verify that the chosen distribution is consistent with the sample data.

Group Interval is a subrange of a distribution that allows similar values to be grouped together and given a frequency count.

Histogram is a plot of the range of values of a variable into intervals and displays only the count of the observations that fall into each interval. (p. A3-9)

Interquartile Range is the difference between the third quartile (75th percentile) and the first quartile (25th percentile). (p. A3-10)

Kurtosis* is the measure of the degree of peakedness and flatness of a curve. The higher the kurtosis, the closer the points of the curve lie to the mode of the curve. A normal distribution curve has a kurtosis of 3. (p. A3-7)

Lognormal Distribution is the distribution of a variable whose logarithm is normally distributed. (p. A3-15)

Mean is the arithmetic average of a set of numerical observations: the sum of the observations divided by the number of observations (p. A3-7).

Measurement Error is error introduced through imperfections in measurement techniques or equipment.

Median is the value midway (in terms of order) between the smallest possible value and the largest possible value. It is that value above which and below which half the population lies (p. A3-7).

Mode* is that value which, if it exists, occurs most often in a data set. (p. A3-7)

Monte Carlo Analysis (Monte Carlo Simulation) is a computer-based method of analysis developed in the 1940's that uses statistical sampling techniques in obtaining a probabilistic approximation to the solution of a mathematical equation or model. It is a method of calculating the probability of an event using values, randomly selected from sets of data repeating the process many times, and deriving the probability from the distributions of the aggregated data.

Non-parametric Approach is one that does not depend for its validity upon the data being drawn from a specific distribution, such as the normal or lognormal. A distribution-free technique.

Normal Distribution is a probability distribution for a set of variable data represented by a bell shaped curve symmetrical about the mean. (p. A3-14)

Parameter. Two distinct, but often confusing, definitions for parameter are used. In the first usage (preferred), parameters refers to the constants characterizing the probability density function or cumulative distribution function of a random variable. For example, if the random variable W is known to be normally distributed with mean μ and standard deviation σ, the characterizing constants μ and σ are called parameters. In the second usage, parameters are defined as the constants and independent variables which define a mathematical equation or model. For example, in the equation $Z=\alpha X+\beta Y$, the independent variables (X,Y) and the constants (α,β) are all parameters.

Parametric Approach is a method of probabilistic analysis in which defined analytic probability distributions are used to represent the random variables, and mathematical techniques (e.g., calculus) are used to get the resultant distribution for a function of these random variables.

Percentile is the value that exceeds X percent of the observations.

Population is the total collection of observations that is of interest.

Probability (Classical Theory) is the likelihood of an event.

Probabilistic Approach is an approach which uses a group of possible values for each variable to estimate risk.

Probabilistic Model is a system whose output is a distribution of possible values.

Probability Density Function (PDF) is a distribution of values for a random variable, each value having a specific probability of occurrence. It is alternatively referred to in the literature as the probability function or the frequency function. For continuous random variables, that is, the random variables which can assume any value within some defined range (either finite or infinite), the probability density function of a point expresses the probability that the random variable falls within some very small interval; the PDF at a point multiplied by the width of a very small interval containing the point approximates the probability that the random variable falls within that interval. For discrete random variables, that is, random variables which can only assume certain isolated or fixed values, the term probability mass function (PMF) is preferred over the term probability density function. PMF expresses the probability that the random variable takes on a specific value.

Quantile-Quantile (Q-Q) Plot portrays the quantiles (percentiles divided by 100) of the sample data against the quantiles of another data set or theoretical distribution (e.g., normal distribution). By comparing the data to a theoretical distribution with a straight line, departures from the distribution are more easily perceived. (p. A3-24)

Random Error is error caused by making inferences from a limited database.

Random Number Generator* is a method implemented in a computer program that is capable of producing a series of independent, random numbers.

Random Variable is a quantity which can take on any number of values but whose exact value cannot be known before a direct observation is made. For example, the outcome of the toss of a pair of dice is a random variable, as is the height or weight of a person selected at random from the New York City phone book.

Range* is the difference between the largest and smallest values in a data set. Alternatively, it expresses the interval between the minimum and maximum values (i.e., (min x_i, max x_i))

Regression Analysis (Simple) is the derivation of an equation which can be used to estimate the unknown value of one variable on the basis of the known value of the other variable.

Sampling. One of two sampling schemes are generally employed: simple random sampling or Latin Hypercube sampling. Latin hypercube sampling may be viewed as a stratified sampling scheme designed to ensure that the upper or lower ends of the distributions used in the analysis are well represented. Latin hypercube sampling is considered to be more efficient than simple random sampling, that is, it requires fewer simulations to produce the same level of precision. Latin hypercube sampling is generally recommended over simple random sampling when the model is complex or when time and resource constraints are an issue.

Sensitivity Analysis is an analysis that attempts to provide a ranking of the model's input parameters with respect to their contribution to model output variability or uncertainty. In broader sense, sensitivity can refer to how conclusions may change if models, data, or assessment assumptions are changed.

The difficulty of a sensitivity analysis increases when the underlying model is nonlinear, nonmonotonic or when the input parameters range over several orders of magnitude.

Simple Random Sampling (SRS) is a sampling procedure by which each possible member of the population is equally likely to be the one selected.

Simulation, in the context of Monte Carlo analysis, is the process of approximating the output of a model through repetitive random application of a model.

Skewness is the measure of the degree of deviation of a curve from the norm of a symmetric distribution. The greater the degree of skewness, the more points of a curve lie to one side of the peak of the curve. a normal distribution curve having no skewness is symmetrical, that is to say that there exists a central value a

such that f(x-a)=f(a-x), f(x) being the frequency function. (p. A3-7)

Standard Deviation is a measure of dispersion which is expressed in the same units as the measurements. It is a measurement of the variability of a distribution, i.e., the dispersion of values around the mean. Standard deviation is the square root of the variance for a distribution (p. A3-7).

Standard Error of the Mean is the standard deviation of the distribution of possible sample means. This statistic gives one indication of how precise the simulation is.

Stochastic is a term referring to a process involving a random variable.

Triangular Distribution is a distribution with a triangular shape. It is characterized by its minimum, maximum and mode (most likely) values. It is often used to represent a truncated log-normal or normal distribution if there is little information available on the parameter being modeled. (p. A3-14)

Variability refers to observed differences attributable to true heterogeneity or diversity in a population or exposure parameter which cannot be reduced by additional data collection.

Sources of variability are the result of natural random processes and stem from environmental, lifestyle , and genetic differences among humans. Examples include human physiological variation (e.g., natural variation in bodyweight, height, breathing rates, drinking water intake rates), weather variability, variation in soil types and differences in contaminant concentrations in the environment. Variability is usually not reducible by further measurement or study (but can be better characterized).

Variance is a measure of the dispersion, or spread, of a set of values about a mean. Variance is the square of the standard deviation, i.e., the average of the squares of the deviations of a number of observations from their mean value. When values are close to the mean, the variance is small. When values are widely scattered about the mean, the variance is larger.

* from Decisioneering manual (see bibliography)

Bibliography

(1997) Air Force Technical Report on Methods to Quantify Uncertainty in Human Health Risk Assessment - DRAFT, Armstrong Laboratory Occupational and Environmental Health Directorate, Brooks Air Force Base, Texas.

Decisioneering, Inc. (1996) Crystal Ball Version 4.0 User Manual, pages 269-275.

Marriott, F.H.C. (1990), a Dictionary of Statistical Terms -Fifth Edition, Longman Scientific and Technical copublished with John Wiley & Sons: 605 Third Avenue, New York, New York 10158, page 9.

U.S. EPA,(October 4,1996) Guiding Principles for Monte Carlo Analysis -DRAFT, U.S.EPA, 401 M Street SW, Washington, DC 20460.

ATTACHMENT 2

ATTACHMENT 2: Probabilistic Risk Assessments and Monte-Carlo Methods: A Brief Introduction

Risk assessments are a crucial part of EPA's pesticide regulatory program and have been for over 25 years. These assessments are used to estimate impacts on human health and the environment from the use of a given pesticide. Agency policy is that risk assessment should be conducted in a tiered approach, proceeding from simple to more complex analyses as the risk management situation requires. The Agency has traditionally used "deterministic" assessments involving point estimates of specific parameters to generate a single estimate of exposure and risk based on various assumptions about the concentration of pesticide in any given medium (e.g., food, water, air etc) and the amount of that medium consumed, breathed, or otherwise contacted. Deterministic assessments can begin with worst-case assumptions (for example, residues on foods at tolerance levels), then can be refined by more realistic values that remain point estimates (for example, average residues from field trials). Even with a tiered approach, each deterministic assessment provides single values for estimates of exposure from a given pathway. Such single-value risk estimates do not provide information on the variability and uncertainty that may be associated with an estimate.

Current Agency Policy (5/15/97) is that *probabilistic* analysis techniques (of which Monte-Carlo is one example) can be viable statistical tools for analyzing variability and uncertainty in risk assessments, provided they are supported by adequate data and credible assumptions. Probabilistic techniques can enhance risk estimates by more fully incorporating available information concerning the *range* of possible values that an input variable could take, and weight these values by their *probability* of occurrence. As an example, a particular food commodity (e.g., tomatoes) might contain a range of pesticide residues for any given pesticide, with a large percentage of tomatoes consumed actually containing no residues at all (since not all tomatoes are treated). In addition, individuals may or may not consume tomatoes on any given day and, over time, are expected to consume varying amounts of this food item due to varying daily consumption patterns. Probabilistic risk analysis permits OPP to assess the range of exposures (and their associated probabilities) which result from combinations of the various residue levels and consumption patterns. The resulting output of a probabilistic determination is a distribution of risk values with probability assigned to each estimated risk. Some of the major differences between deterministic and probabilistic estimates are summarized in the table below:

Deterministic Risk Assessment	Probabilistic Risk Assessment

• Pesticide concentrations and potential exposure factors are expressed as point estimates. • The risk estimate is also expressed as point value. The variability and uncertainty of this value are not reflected.	• Takes into account all available information and considers the *probability* of an occurrence. • The risk estimate is expressed as a distribution of values, with a probability assigned to each value. • The distribution reflects variability and uncertainty.

Tiered Approach to Risk Assessment

As risk assessments are refined, assumptions can proceed from more conservative (more health protective) to more realistic reflections of exposure. As noted above with the example of residues on food, such refinements can be applied to deterministic assessments. Probabilistic analyses, including Monte Carlo, represent numerical techniques to reflect more realistic assumptions. For example, Tier I of acute dietary assessments as conducted by OPP includes conservative assumptions such as: all foods consumed by an individual in any given day were treated with the pesticide in question (if registered for use on that food) and that residues are present in those consumed foods at the maximum legal limit. Monte-Carlo techniques fully applied to this situation would allow incorporation of information concerning the percent of the crop which is treated, the amount of pesticide applied and timing of its application, and the range and distribution of residue values expected to be found. This information is useful because a particular food (e.g., tomatoes) might contain a range of pesticide residues for any given pesticide, with a large percentage of tomatoes consumed actually containing no residues at all (since not all tomatoes are treated). In addition, individuals may or may not consume tomatoes on any given day and, over time, are expected to consume varying amounts of this food item due to varying daily consumption patterns. Any variability and uncertainty is explicitly included in the analysis and is fully disclosed.

The Origin of Monte-Carlo Techniques

Monte-Carlo techniques have been used since the 1940's when they were first developed by physicists working on the Manhattan project. Only recently, however, have personal computers become sufficiently powerful and widespread for Monte-Carlo techniques to be widely applied for health risk assessments. Modern spreadsheet programs now provide a range of critical facilities to help to illustrate and order a model including advanced statistical functions, charting, etc. And the simplicity and capabilities of recently introduced commercial Monte-Carlo software allows these techniques to become virtually all but routine.

The origin of the name “Monte-Carlo” relates to the famous gambling city in Monaco, but the relation to gambling applies only to the probability of a given event occurring over the long term. Although one cannot

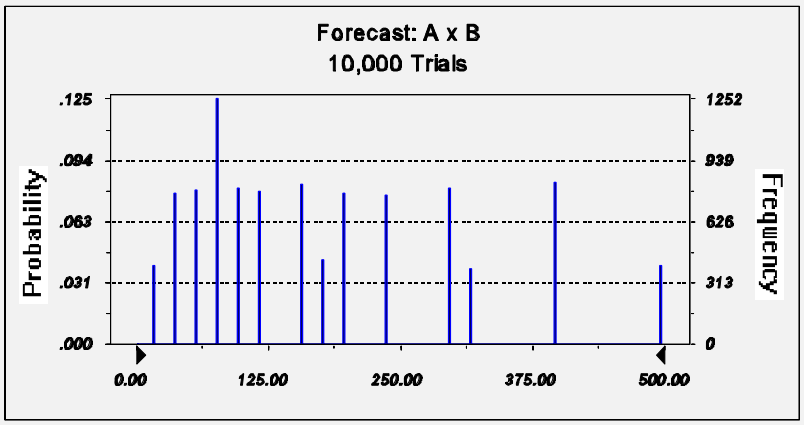
know precisely which number will appear on the next roll of a craps die or the spin of a roulette wheel, one can predict over the long term (and as precisely as desired) the frequencies associated with each outcome. Monte-Carlo numerical techniques similarly cannot predict exactly which exposures will occur on any given day to any specific individual, but can predict the range of potential exposures in a large population and each exposure's associated probability.

What is Monte-Carlo Analysis?

Monte-Carlo analysis is simply one of several mathematical techniques for performing probabilistic risk assessments. The Monte Carlo technique, as applied to exposure assessment, involves combining the results of hundreds or thousands of random samplings of values from input probability distributions in such a manner as to produce an output distribution which reflects the expected range and frequency of exposures. Although computationally-intensive, Monte-Carlo techniques themselves are not complicated. Assessing a Monte-Carlo analysis requires examining the appropriateness of assumptions, judgements, and data sets which are key inputs to the mathematical procedures.

The first step in a Monte-Carlo simulation is the construction of a model that accurately represents the problem at hand. The makeup of the model usually entails a mathematical combination (addition, multiplication, logarithms, etc.) of the model input variables which can be expressed as probability distributions. If, for example, one desires to simulate the daily dietary pesticide exposure to individuals from a particular pesticide in tomatoes, this can be simulated by repeatedly drawing random values from two separate distributions: one distribution represents tomato consumption by individuals while the other represents pesticide levels in tomatoes. Here, the output variable (daily pesticide exposure) is defined as the product of the two input variables (tomato consumption in grams/day and pesticide residue concentrations in ug/g). Each random pair of input variables obtained from repeated independent samplings of the input distribution are multiplied together and the product used as one point in the distribution for the output variable. In general, this process is repeated thousands of times and the thousands of output products generated, taken together, form a distribution of frequencies. This technique is more fully illustrated in the box on the following page:

Suppose that our two input variables are defined as **A** and **B** where **A** = {2 , 4, 6, 8, 10}, **B** = {10, 20, 30, 40, 50}, and our output variable **C** is defined as the product of **A** and **B** (i.e., **C** = **A** x **B**). Set **A** might represent the concentration of a pesticide in tomatoes (in ug/g) and Set **B** might represent the daily consumption of tomatoes (in g/day). We wish to determine the range and frequency of potential values of **C** (which in this case would represent daily exposure to the pesticide in ug/day) . Inspection of the input data immediately reveals that the value for **C** (daily exposure) can range from a low of 20 ug/day (i.e., 2 x 10) to a high of 500 ug/day (i.e., 10 x 50) and that each of these two extreme values should occur approximately 4% of the time (i.e., 1/5 x 1/5 = 1/25 = 4%), Monte-Carlo methods permit us to evaluate **all** values that can be generated for the value **C** along with each of their associated probabilities. The Monte-Carlo method randomly chooses a single pesticide concentration value from Set **A** and a single tomato consumption value from Set **B**. These two values are multiplied together (to give daily pesticide exposure, **C**) and this resultant value stored. This process is repeated thousands of times with all values of **C** eventually plotted as a frequency histogram as shown above. Note that the lowest value is 20 ug/day and the highest value is 500 ug/day, just as originally predicted. Note also that these two values each occur approximately 4% of the time, just as (again) predicted from our original inspection. Although this example uses discrete values for sets **A** and **B**, Monte-Carlo modeling can also be performed when the input variable are described as continuous variables.



Regardless of how accurately the fitted distribution conforms to the data, or what method of sampling is used, the analyst has to set up a model that reflects the situation being assessed. According to Vose's *Quantitative Risk Analysis: A Guide to Monte Carlo Simulation Modeling*, the cardinal rule of risk analysis modeling is: "Every iteration of a risk analysis model must represent a scenario that could physically occur." Following this rule will lead to a model that is both accurate and realistic. As an example, it would be improper to model a cow diet as a random sampling of feeds with established tolerance for the pesticide of interest since many of the diets generated in such a manner would be unreasonable with respect to the roughage/nonroughage components, carbohydrate/protein mix, commodity combinations, and economic constraints. In short, blind application of Monte-Carlo techniques without regard for the reality of the generated scenarios will produce absurd results with no basis in reality. The analyst should ensure that each of the hundreds or thousands of iterations is a scenario with real-world possibilities.

It is often tempting in risk analysis modeling to include very unlikely events that would have a very large impact should they occur. A rare event of concern is defined as an event that has a low probability of

occurrence but a potentially high impact on the results of a risk analysis. The expected impact of a rare event is determined by two factors: the probability that it will occur and the distribution of possible impacts. For example, widespread systematic illegal use of a pesticide or gross calibration errors in a pesticide's application might be a situation which occurs to some unknown (but relatively insignificant) extent. Since the probabilities of these events are so difficult to quantify, their determination provides a stumbling block for the analyst. However, since it is impossible to cover all scenarios that might exist and to calculate the probability of such occurrence, including the rare event in the general model will not increase our understanding of reality and will limit the clarity of the model.

Random Nature of the Monte Carlo Analysis.

Integral to any Monte-Carlo analysis is the generation of random numbers. Similar to rolling dice, the software has a 'random number generator' that produces a random sequence of numbers. Two main forms of sampling are Random Sampling (also called Monte Carlo Sampling) and Latin Hypercube sampling. Random or Monte Carlo sampling will evaluate the probability distributions in a purely random fashion, and is useful in trying to get the model to imitate a random sampling from a population or for doing statistical experiments. However, the randomness of this sampling suggests that, unless a very large number of iterations are performed, it will over-sample some parts and under-sample other parts of the distributions. Because for nearly all risk analysis modeling exploration of the distribution extremes (the "tails") is important, exact reproduction of the contributing distributions of the model becomes essential.

Latin Hypercube sampling (LHS) addresses this issue by providing a sampling method that appears random but that also guarantees to reproduce the input distribution with much greater efficiency that the random sampling. LHS uses a technique known as stratified sampling without replacement. It breaks the probability distribution into 'n' intervals of equal probability, where 'n' is the number of iterations to be performed on the model. Then, at random, one sample is drawn from each section, forcing, this way, an equal-chance representation of all the portions of the distribution. The Latin Hypercube method leads to a predictable uniformity of the sampling of the distribution.

For More Information

Use of Probabilistic Techniques (Including Monte Carlo Analysis) in Risk Assessment, Memorandum from the Office of the Administrator, U.S. Environmental Protection Agency, May 15, 1997

Policy for Use of Probabilistic Analysis in Risk Assessment at the U.S. Environmental Protection Agency. U.S. EPA, Office of Research and Development, May 15, 1997. (http://www.epa.gov/ncea/mcpolicy.htm)

Vose, David. Quantitative Risk *Quantitative Risk Assessment: a Guide to Monte-Carlo Simulation Modeling*. John Wiley and Sons (1996)

ATTACHMENT 3: Distribution Selection

Section I Introduction

EPA has recently established a policy and a series of guiding principles for the use of various probabilistic risk assessment techniques. The policy states that probabilistic risk analysis techniques (including Monte-Carlo analyses) can be viable statistical tools for analyzing variability and uncertainty in risk assessments provided that adequate supporting data are available and credible assumptions are made. The policy goes on to state that when risk assessments using probabilistic techniques are submitted to the Agency for review and evaluation, a number of conditions must be satisfied: these conditions relate to the good scientific practices of transparency, reproducibility, and the use of sound methods (memo from F. Hansen, 5/15/97). One of these specific conditions of acceptance states that

> The methods used for the analysis (including all models used, all data upon which the assessment is based, and all assumptions that have a significant impact upon the results) are to be documented and easily located in the report. This documentation is to include a discussion of the degree to which the data used are representative of the population under study. Also, this documentation is to include the names of the models and software used to generate the analysis. Sufficient information is to be provided to allow the results of the analysis to be independently reproduced.

The Agency simultaneously released a series of sixteen "Guiding Principles" for the use of Monte-Carlo analysis and an Appendix dealing with the selection of appropriate input probability distributions for these analyses. The intent of the current document is to further develop these principles and guidelines for use by pesticide registrants and other interested parties by defining what we in OPP's Health Effects Division (HED) see as key criteria which a risk assessments using Monte-Carlo risk assessment techniques must adequately address. Specifically, this chapter explores the various plots, tests, techniques, and analyses which could be used to define an adequate probability distribution for use as an input parameter for a Monte-Carlo assessment submitted to HED.

Monte-Carlo Modeling Options

Once the raw input data on the exposure variable of interest is collected, a risk assessor has available a number of techniques for representing the exposure variables in a Monte Carlo analysis.

- an assessor can use the data values themselves directly in the simulation in what is termed a "trace-driven" simulation. In this technique, values from the raw input data are repeatedly selected in a random manner and used to calculated model outputs; This is one form of an empirical distribution function (EDF) in which the data values (and only the data values) themselves are used as model inputs.

- an assessor can use the data to define a non-parametric empirical distribution function (EDF) where the data values themselves are used to specify a cumulative distribution and the entire *range* of values (including intermediate points) is used as model inputs. With this technique, *any* value between the minimum and maximum observed values can be selected and model input is not limited to the specific values present in the measured data.

- an assessor can attempt to fit a theoretical or parametric distribution (PDF) to the data using standard statistical techniques and input parameters to the model can be selected from this fitted distribution.

In evaluating whether an EDF or PDF should be used, the number of field trials, number of residue values, the percent of crop treated, and the visual fit between a theoretical distribution and the actual data are important considerations. A key consideration is the determination of how likely it is that we have captured (or adequately estimated) the high end residues which are present in the population of treated and untreated commodities.

There are a number of potential benefits for making distributional assumptions about exposure data (du Toit *et al*, 1986; Law and Kelton, 1991). For example,

1) Distributional assumptions permit the data to be represented compactly. A data set containing a potentially large amount of information can be summarized as a probability distribution model described by only a few parameters. Empirical distributions require that each data point be represented and can result in a data set that is cumbersome and difficult to use if the data set is large.

2) Distributional assumptions (and the exploratory data analysis which precedes them) may lead to a clearer understanding of the underlying physical mechanisms involved in generating the data and *vice-versa*.

3) Distributional assumptions permit data to be generated which are *outside* the range of historically observed data. This can be useful since many measures of performance for simulated systems depend heavily on the probability of an "extreme event" (i.e., one outside the range of the observed data) occurring. Empirical distributions, which rely solely on past data when used in the usual manner, can tend to underestimate the probability of an extreme event.

4) Distributional assumptions permit the data to be "smoothed out" which may more accurately reflect real-world values. Empirical distributions, on the other hand, may contain certain artifactual irregularities, particularly if only a small number of data values are available.

On the other hand, some authors prefer EDFs (Bratley, Fox and Schrage, 1987) arguing that the smoothing which necessarily takes place in the fitting process distorts real information. In addition, when data are limited, accurate estimation of the upper end (tail) is difficult. Unfortunately for the assessor, there is no consensus as to which method is best. In general, the use of parametric (theoretical) distributions may be preferable to the use of empirical distributions when the data are limited, the fit of the theoretical distribution to the data is good, and there is a theoretical or mechanistic basis which supports the chosen parametric distribution. The process of selecting probability distributions and evaluating the goodness-of-fit is a process that requires judgement. Ultimately, the technique selected will be a matter of the quality and quantity of the data under evaluation and the assessor's exercise of intelligence, creativity, and honesty in assessing the variability and uncertainties inherent in the risk assessment problem.

Organization of Document

Section I of this document is this introduction to Monte-Carlo methods and a brief description of the advantages of disadvantages of parametric methods (i.e., methods which make assumptions about underlying distributions to develop theoretical distributions) and non-parametric methods (which utilize the data directly in forming an empirical distribution, thereby making no assumptions about underlying distributions).

Section II of this document focuses on parametric methods for characterizing and quantifying stochastic variability. In this section, it is explicitly assumed that the risk assessor has previously made the judgement that the data in hand

are of acceptable quality and are acceptably representative of the exposure variable of interest. The discussion in this parallels the Guiding Principles section and Technical Appendix of the Agency's policy for Monte Carlo Analysis, expanding these elements to provide more technical detail. The general outline in Section II follows that developed by Law and Kelton (1991). It is organized around three fundamental activities:

(*I*) *selecting candidate theoretical distributions* to determine which general families appear to be appropriate to use on the basis of the shape, summary statistics, and simple distributional plots;

(*II*) *estimating the intrinsic parameters of the candidate distributions* to define the specific distribution; and

(*III*) *assessing the quality of the resulting fit* by examining how closely they represent the true underlying distributions for the data of interest and using various Goodness-of-Fit (GoF) tests.

Assessors have a wide variety of commercially available distribution-fitting programs, spreadsheets, and dedicated statistical packages to assist them in deciding whether or not their data can be adequately represented by a theoretical distribution function. It is expected that most assessors will make use of one or more of these programs in fitting exposure data. While these programs can save a tremendous amount of work, their use should never be reduced to a simple mechanical exercise of importing the data, running the analysis and picking the "best fitting" distribution returned by the program. Furthermore, despite their obvious utility, many of the commercial fitting-packages are limited for fitting exposure data. For example, most fitting packages currently available cannot fit singly or multiply censored data, truncated distributions, or distributional mixtures. For these data, the assessor will have to seek more selective, powerful tools.

Many times in Monte Carlo analyses, an empirical distribution function (EDF) is used to characterize a model variable if the risk assessor has determined that the data themselves provides the best representation of the exposure variable. In Section III, we define an EDF and discuss the conditions under which the use of an EDF may be preferable to a CDF. Several approaches used to implement EDFs are also discussed.

Throughout Sections II and III, each key idea will be illustrated through a case study example.

Section II Parametric Methods

Parametric methods (as opposed to the non-parametric or empirical methods discusses in Section III) rely on a mathematical description of the *distribution* of values generated by a process. This section of the document describes the three standard activities (selecting candidate distributions, estimation of parameters, and assessing goodness-of-fit) used to describe the distribution and the adequacy of this description. The general outline follows that developed by Law and Kelton (1991).

Activity I – Selecting Candidate Distributions

Activity I involves the use of prior knowledge and exploratory data analysis to make preliminary assessments of which general *families* of distributions appear to best match the input data. This evaluation is performed on the basis of the shape, summary statistics, and simple distributional and graphical plots of the input data and does not, at this stage, involve the estimation of the specific statistical parameter values associated with each of these families.

Knowledge of the various properties and parameters associated with any of the various potential distributions can aid in the selection of an appropriate distributional family. Figure 1 provides a flow chart which may be used as a guide to selecting potential distributions for further analysis based on prior knowledge of distribution characteristics. It is not intended to be all-inclusive, but does cover a range of distributions which might be commonly seen in the area of exposure and health risk assessment.

Make Use of Prior Knowledge

The choice of input distribution should always be based on all relevant information (both qualitative and quantitative) available for a parameter. In selecting a distributional form, the risk assessor should consider the quality of the information in the database and ask a series of broad questions which might include the following:

Is there any mechanistic basis for choosing a distributional family? Is the shape of the distribution likely to be dictated by physical or biological properties or other mechanisms? Ideally, the selection of candidate probability distributions should be based on consideration of the underlying physical processes or mechanisms thought to be key in giving rise to the observed variability. For example, assume that a persistent systemic pesticide is present in a lettuce plant and is not degraded or metabolized. If, due to weekly variations in sunlight, rainfall, and nutrient availability, the mass of each lettuce leaf increases each week by some random independent proportion of the mass achieved during the previous week, the distribution of residues in these lettuce plants will be lognormally distributed (Ott, 1995); in this case, the residue concentrations can be expressed as a random proportion of the quantity present in the immediately prior state. If each successive proportion is independent of the one before and many weeks pass between the initial and final states, the final residue concentration in the lettuce plant can be expressed as a product of random variables which gives rise naturally to a lognormal distribution. In general, if an exposure variable is the result of the product of a large number of other random variables, it would make sense to select a lognormal distribution for testing. As another example, the exponential distribution would be a reasonable candidate if the stochastic variable represents a process akin to inter-arrival times of events that occur at independent constant rates.

Is the variable discrete or continuous? Can the variable only take on discrete values or is the variable continuous over some range? A discrete variable may only take one of several specific values, whereas a continuous variable may take on an infinite number of values. Examples of discrete variables would include

whether the crop is treated or not (e.g., 0 or 1), the number of times a given pesticide is applied per season, or the number of showers taken per week. Examples of continuous variable include the residue concentration of a given pesticide in a tomato, the amount of pesticide a.i. applied per acre in a season, or drinking water consumption rate.

Is the variable bounded or unbounded? If bounded, what are the bounds of the variable? What is the physical or plausible range of the variable? Is it semi-infinite (X>b)? Does it take on only positive values (X>0)?; Is it bounded by the interval [a,b]? A properly-fitted distribution should cover the range of values over which the modeled variable could theoretically extend. If a fitted distribution extends beyond the range of plausible values, then the model will produce implausible scenarios at the extreme tails of the distribution. Conversely, if a fitted distribution fails to adequately extend to cover real-world limits, the resulting model will not reflect the true nature of the potential variability.

Beta distributions are examples of bounded continuous distributions which might be considered for percent foliar dislodgeable residue (%FDR) which could vary between 0% and 100%, for example. Unbounded continuous distributions include the normal distribution: these distributions can sometimes be truncated, if necessary, to represent variables which have natural or practical physical limits (e.g., body weight). Semi-infinite continuous distributions (X>0) include the exponential distribution, the gamma distribution, the log-normal distribution, and the Weibull distribution. These distributions are all bounded on one-side (sometimes by 0) and extend to infinity and may describe variables which are censored due to limits of detection or some aspect of the experimental design. It is important to note that a correctly fitted distribution can extend *beyond* the range of observed data. This is expected since data are rarely observed at the theoretical extremes for the variable in question.

Are historical data available? Is it known that a variable of interest has been found to consistently have a certain distribution type in other data collection and distribution fitting research? Previous data may be available for similar (or even identical) situations. For example, environmental concentrations of a contaminant have sometimes to be found to be lognormally distributed. Time to complete certain tasks have been shown to follow in some cases a Weibull distribution. Human body weights have been modeled as a normal or log-normal distribution (Burmaster and Crouch, 1997). Consumption of water have been shown in some instances to be adequately represented by a log-normal distribution (see, e.g., EPA's *Exposure Factors Handbook*, the AIHC's *Exposure Factors Sourcebook*), or Roseberry and Burmaster (1997). A registrant should be aware of past modeling attempts to incorporate distributional information and may wish to incorporate this into its own assessments.

Does the sample represent a single population, or is the sample drawn from a mixture of subpopulations? Mixture models arise frequently in exposure and risk assessment. Discrete mixture distributions occur when the population of interest consists of a number of distinct subgroups, each with their own unique distribution. For example, different agricultural occupational groups may have different exposure distributions as a result of differing activities; produce grown in different regions of the country may have systematic differences in pesticide residue concentrations due to systematic differences across the U.S. in rainfall and rainfall patterns, soil types and conditions, and length of the growing season . Multi-modality provides a first strong suggestion that the observed sample is drawn from a mixture of distributions and is therefore not homogenous. As a second step, statistical tests (e.g., the non-parametric Kruskal-Wallis test) are available for assessing the homogeneity of different data sets (e.g., Florida residue data vs. California residue data) and determining whether the data sets can indeed be merged into *the* single residue distribution. Distinguishing between these different subgroups can be important for both scientific evaluations of risk and evaluations of different distributional issues. When these differences are recognized and the subgroups identified, the overall distribution can be built up from the individual distributions of the various subgroups.

Explore the Data

Exploring the data is an important step in the process of selecting plausible distributions. Exploratory data analysis can be thought of as consisting of two steps: (1) *characterizing the data through the use of summary statistics* and (2) *graphical data analysis*.

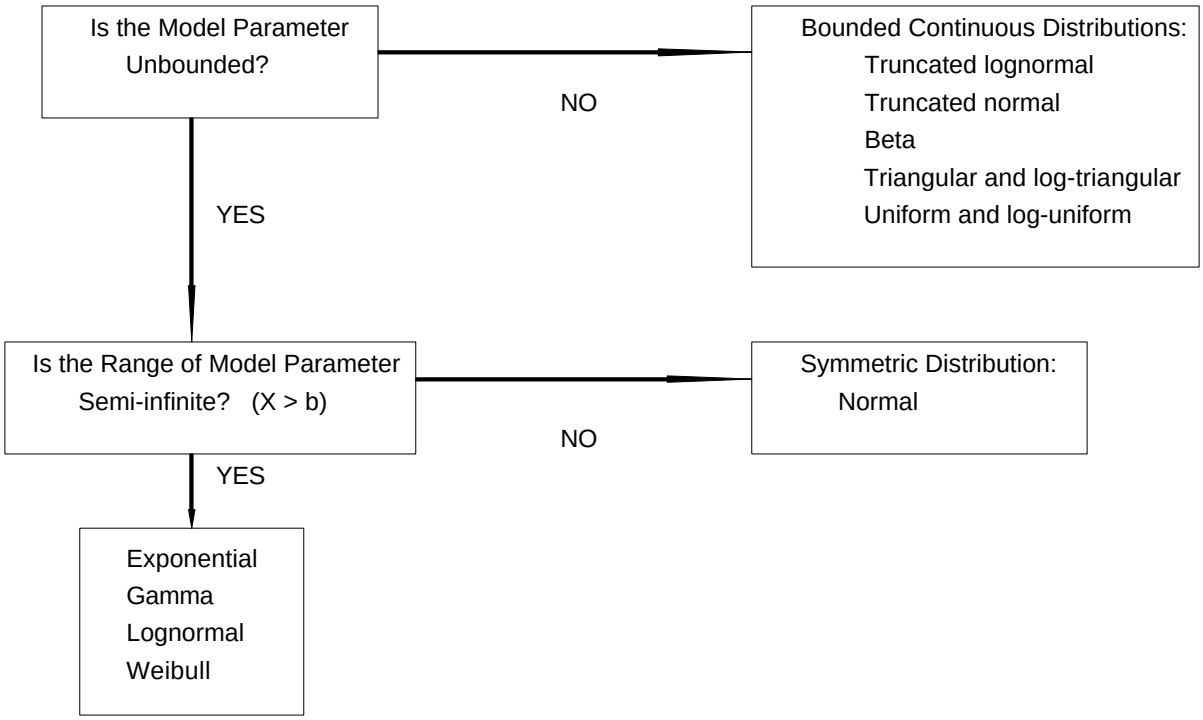
Summary Statistics. Summary statistics are useful for initially characterizing or describing the data. Common summary statistics fall into three basic groupings: (1) *measures of central tendency or location*, such as the mean or median; (2) *measures of dispersion or spread*, such as the variance; and (3) *measures of shape or skewness*.

Measures of central tendency are intended to indicate the "center" of the data and commonly include the mode, median, and mean. Other measures of location include the geometric mean and trimmed mean (Helsel and Hirsh, 1992).

Measures of spread are intended to indicate how dispersed the data are relative to some central value or specify the distance between selected observations. Common measures of spread include the range, inter-percentile ranges (e.g., inter-quartile range), standard deviation, variance, and coefficient of variation.

Measures of shape are intended to provide insights to the symmetry or asymmetry in the distribution of the data. The most frequently used measures of shape are skewness (asymmetry) and kurtosis (degree of peakedness). In some cases, these summary statistics can be used to suggest one or more appropriate distribution families for further testing as part of Activities II and III. For symmetric continuous distributions such as the normal, the mean and the median are equal. Thus, if the mean and median for any given data set are approximately equal, one might consider further analysis of the data to test the hypothesis that the distribution is normal. For exponential distributions, the coefficient of variation (defined as the standard deviation divided by the arithmetic mean, and sometimes expressed as a percent) is equal to 1 (or 100%). Therefore, if the mean and standard deviation of any given data set are numerically similar, an exponential distribution might be an appropriate distribution to hypothesize. Skewness and kurtosis values, considered together, can be used to assist in distribution selection The skewness value is a measure of the symmetry of the data, with perfectly symmetric distributions (like the normal) having a skewness value of zero. Right-skewed distributions, like the right-skewed lognormal, have positive skewness values whereas left skewed distributions have negative skewness values. Exponential distributions have a skewness value of 2. Thus, if a set of data has a coefficient of variation of approximately 1 and a skewness of approximately 2, an exponential distribution would be appropriate to consider. Many statistical and spreadsheet packages have built-in features for automatically calculating many summary statistics. Simply inspecting these output values can aid substantially in determining
candidate distributions for further analysis.

Figure 1. Selecting Continuous Theoretical Distributions



Box 1 lists data used as the case study throughout this section. The data in this Box represent a set of 25 hypothetical residue values in tomatoes. Several summary statistics for these residue data are shown in Box 2. A quick visual inspection of the data can reveal a number of important insights. Box 3 illustrates some of these insights for the sample tomato pesticide data.

BOX 1: Hypothetical Pesticide Concentrations in Tomatoes (ppm)

110.5	204.3
147.5	148.3
111.6	66.9
139.0	53.6
72.9	68.5
109.8	108.0
94.8	97.6
68.8	78.2
142.3	68.2
70.8	80.3
74.6	267.7
169.7	170.0
143.7	

Graphical Data Analysis. The risk assessor can often gain important insights by using a number of simple graphical techniques to explore the data prior to numerical analysis. The importance of this phase of visual inspection cannot be over-emphasized. A wide variety of graphical methods have been developed to aid in this exploration including frequency histograms, stem and leaf plots, dot plots, line plots for discrete distributions, box and whisker plots, and scatter plots [Tukey (1977); du Toit et al. (1986); Morgan and Henrion, (1990)]. These graphical methods are all intended to permit visual inspection of the density function corresponding to the distribution of the data. They can assist the assessor in examining the data for skewness, behavior in the tails, rounding biases, presence of multi-modal behavior, and data outliers. Graphical methods, however, can be highly misleading in the face of considerable uncertainty due to small sample size or a high coefficient of variation.

BOX 2: Summary Statistics for Hypothetical Pesticide Concentration in Tomatoes (ppm)

Quantiles

maximum	100.0%	267.70
	99.5%	267.70
	97.5%	267.70
	90.0%	183.72
quartile	75.0%	145.60
median	50.0%	108.00
quartile	25.0%	71.86
	10.0%	67.68
	2.5%	53.60
	0.5%	53.60
minimum	0.0%	53.60

Moments

Mean	114.7056
Std Dev	51.2019
Std Error Mean	10.2404
Upper 95% Mean	135.8405
Lower 95% Mean	93.5707
N	25.0000
Sum Weights	25.0000
Sum	2867.6400
Variance	2621.6304
Skewness	1.2857
Kurtosis	1.8846
CV	44.6376

A frequency histogram is a graphical estimate of the empirical probability density function and can be compared to the fundamental shapes associated with standard analytic distributions (e.g., normal, lognormal, gamma, Weibull). Law and Kelton (1991) and Evans et al. (1993) have prepared a useful set of figures which plot many of the standard analytic distributions for a range of parameter values. Frequency histograms can be plotted on both linear and logarithmic scales and should be plotted to avoid too much jaggedness or too much smoothing (i.e., too little or too much data aggregation). If the appearance of the histogram does not change much when varying the bin width over a reasonably wide range, then the data analyst can feel confident that any observed patterns are genuine. If, on the other hand, the appearance changes in a fundamental way depending on the selected bin width, any observed patterns at a specific bin width may be an artifactual and should not be trusted. As a starting point, some authors suggest that it may be useful to select the number of bins according to $k = 1 + 3.322 \log_{10} n$ where n is the number of data points.

Line graphs apply to discrete random variables and are estimates of the probability mass function. In a line graph, the proportion of values in the sample data set equal to a particular

discrete value are plotted and compared, on the basis of shape, to the probability mass functions for discrete distributions (e.g., binomial, geometric, Poisson, negative binomial, etc.).

Box plots (Tukey box plots, box and whisker plots) can be a very effective graphic display for summarizing the distribution of a data set. Box plots provide easily explained and easily comprehended visual summaries of:

- the center of the data (median - the center line of the box)
- the spread in the data (inter-quartile range - the box length)
- the skewness (quartile skew - the relative size of the box halves)
- the range (whiskers - lines from the ends of the box to the maximum and minimum of the data or to some other selected endpoint, e.g., the 5th and 95th percentiles, etc.)

BOX 3: Distributional and Statistical Insights into Hypothetical Tomato Pesticide Data Set

A number of important insights on the data and its distributional form can be gained by inspecting the summary statistics commonly provided by standard statistical packages. If the distribution is normal, for example, the mean will be approximately equal to the median. From the statistics provided in Box 2, we see that the median of 108.0 is located within the 95% confidence interval of the mean (i.e., 93.6 to 135.8). We also see that the coefficient of variation of 0.446 (44.6%, as indicated in the statistical output) is less than 1, indicating that a normal distribution might be appropriate to hypothesize. Since the mean of 114.7 and standard deviation of 51.2 are not equal, an exponential distribution is unlikely to be appropriate. The skewness value of 1 (as opposed to 2) further supports the elimination of the exponential distribution as a viable candidate for further consideration.

There are three basic versions of the box plot: (1) the *simple box plot*, (2) the *standard box plot*, and (3) the *truncated box plot*. In the commonly-used *standard box plot*, the whiskers extend only to the last data point within one step beyond either end of the box. A step is defined as 1.5 times the length of the box or approximately 1.5 times the inter-quartile range. Data points beyond 1.5 steps of either end of the box are plotted as individual points. When constructed in this manner, the box plot provides a rapid visual impression of the prominent features of the data. The median (or central line within the box) shows the location of the center of the data. The spread of the central 50% of the data are represented by the length of the box. And the length of the whiskers (relative to the box) show how stretched the tails of the distribution are. Individual points which extend beyond the whiskers are outside values which may be further investigated and provide clues as to the distributional form. If the distribution is symmetric (e.g., as with a normal distribution), the box will be divided into two equal halves by the median, the upper and low end whiskers will be the same length, and the number of extreme data points will be distributed equally on either end of the plot Two other kinds of box plots (simple and truncated box plots) are more fully discussed by Helsel and Hirsh (1992).

Because of the variety of box plots available, the potential for confusion exists and all box plots submitted to HED should be clearly labeled as to which values are being represented.

Formal Tests for Normality and Lognormality

While examination of the summary statistics, frequency histograms, and box-and-whisker plots associated with a data set are useful exercises in exploratory data analysis, several procedures are available to formally test for normality (or lognormality when log-transformed data are used) and can be used to confirm the assumption of normality/lognormality. Such tests include Shapiro-Wilks test (for sample sizes $\leq$ 50), D'Agostino's test (for sample sizes 50), and Filliben's statistic (sample size >50), which is an extension of the Shapiro-Wilk test. The Shapiro-Wilk and D'Agostino tests are the tests of choice when testing for normality (or lognormality) and are more fully described in a number of standard texts. While the Shapiro Wilk test is one of the most powerful tests for normality, it is difficult to implement by hand as it involves calculating a correlation between the quantiles of the standard normal distribution and the ordered values of the data set. It is, however, easily implemented as part of many

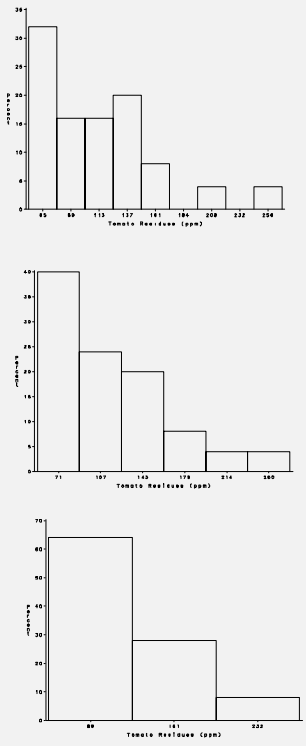
statistical software packages. These tests (and many more) are more fully discussed in the EPA publication *Practical Methods for Data Analysis* (U.S. EPA, 1996). This EPA publication is available on-line and can be downloaded in PDF format (see References and Suggested Readings for http:// address)

It is important to remember during this activity that it is less critical for the analyst to be able to state with absolute certainty that the data are distributed in the hypothesized manner (e.g., lognormally) than it is to determine that the hypothesized distribution is "adequately representative" of the data. The basic question to be answered in the affirmative is whether the empirical distribution of the data is sufficiently well-approximated by the hypothesized distribution for the intended purpose.

Knowledge of the various properties and parameters associated with any of the various potential distributions can aid in the selection of an appropriate distributional family. A list of selected theoretical distributions is included in Table 1 along with a brief description of some of their potential uses. As with Figure 1, it is not intended to be all-inclusive, but does cover a range of distributions which might be commonly seen in the area of exposure and health risk assessment.

BOX 4: Frequency Distribution Histograms for Hypothetical Pesticide Data

For a histogram of the pesticide residue data, the initial number of number of bins is estimated as $k = 1 + 3.322 \log_{10}(25)$ 6 The figures below show histograms for the tomato residue data for 3, 6, and 9 bins. For these data, 6 bins appear to strike a reasonable balance between too much smoothing for the 3 bin histograms and too much jaggedness apparent for the 9 bin histogram.



BOX 5: Determination of the Appropriate Distributional Family for the Hypothetical Residue Data

Box 4 suggested that a normal distribution would be appropriate to hypothesize for the hypothetical pesticide data. However, the box and whiskers plot of the actual data reveals a decidedly right-skewed distribution; in addition the Shapiro-Wilk statistic of 0.88 ($p<0.0063$) also suggests that a normal distribution is not appropriate. As indicated before (and confirmed by the shape of the histogram and box-and-whisker plot), an exponential distribution is also inappropriate for further consideration. Log-transformation of the hypothetical data produces a symmetric mound-shaped histogram and a box-and-whisker plot showing characteristics of the normal distribution (eg., a box divided into two equal halves by the median, whiskers of similar length, and an equal number of extreme data points on either end of the plot). The summary statistics further suggest that a lognormal distribution may be appropriate (mean median and a skewness value substantially closer to 0); the Shapiro-Wilk test ($W = 0.951$ with $p = 0.27$) confirms this as an appropriate distribution for further consideration and analysis as part of Activity II.

Having determined that the log-normal distribution is the distribution most appropriate for further analysis, the two subsequent activities are determining the most appropriate distribution (Activity II) and performing tests to verify that the selected distribution and its parameters adequately fit the empirical data.

Table 1 Selected Theoretical Distributions[a]

Distribution Type	Distribution Description
Discrete	
Bernoulli	The Bernoulli distribution is used to model random events when there are only two possible outcomes (e.g., success or failure, treatment or no treatment) and is used to generate other discrete random variables (e.g., binomial, geometric, and negative binomial). A Bernoulli random variable can be thought of as the outcome of an experiment that either "fails" or "succeeds" and is fully characterized by its parameter *p*, representing the probability of an event occurring.
binomial	The binomial distribution models the number of successes in *n* independent Bernoulli trials, with the with probability *p* of success in each trial. It is produced by processes that (1) can produce only one or the other of two outcomes and (2) are carried out a finite number of trials. It is fully characterized by the parameters *n*, *p*, and *x* representing the number of trials, the probability of success in each trial, and the number of successes, respectively.
discrete uniform	The discrete uniform distribution models random occurrences when there are several possible outcomes, each outcome with the same probability of occurrence. This is typically used as a "first" model for a quantity that is varying among integers, but about which little is known.
geometric	The geometric distribution models the number of failures before the first success in *n* independent Bernoulli trials, each trial with an identical probability of success. It is a direct analogue of the exponential model except is limited to integers.
	.
Bounded Continuous	
beta	The beta distribution is a very flexible distribution capable of exhibiting a wide variety of shapes. It is often used to model bounded data, to model distributions for proportions or fractions, or to model time to complete some task. It can also be used as a rough model in the absence of data (see Law and Kelton, 1991). Two parameters suffice to describe this distribution (α_1, α_2)
triangular, log-triangular	The triangular distribution is often used a rough model in the absence of data when the values toward the middle of a range of possible values are more likely to occur than values near either extreme. There is no mechanistic basis for this model which is typically used to represent subjective uncertainties. If the range covers several orders of magnitude, the log-triangular distribution is sometimes used. The minimum, maximum, and most likely value suffice to describe this distribution.
uniform, log-uniform	The uniform distribution is often used in the absence of data as a crude model when the quantity is known to randomly vary between known limits but where little else is known. Its use is appropriate when we are able to identify a range of possible values, but are unable to determine which values within the range are more likely to occur than others. The minimum and maximum values suffice to describe this distribution. If the limits cover several orders of magnitude, the log-uniform is sometimes used.

Unbounded Continuous	
normal	The normal distribution models phenomena that are the result of the sum of many other random variables (by the Central Limit Theorem). In other words, if a large number of variables are added together (such that no one variable contributes a substantial amount to total variation), the result will take the shape of a normal distribution. These frequently involve small measurement errors of various types and any process whose final outcome is the result of many independently determined sums. The mean and standard deviation suffice to describe this distribution. The skewness of the normal distribution is 0 (it is symmetric) and the kurtosis is 3. As negative quantities can be generated with the normal distribution, this is in some cases theoretically inappropriate. However, as long as the coefficient of variation of less than ca. 0.2, generation of negative values is sufficiently improbable so as not to be of concern since the probability of generation of values more than five standard deviations from the mean is quite small.

Non-negative Continuous	
exponential	When events are purely random, the times between successive events are described by an exponential distribution. The exponential distribution is frequently used to describe the time between events for Poisson processes (i.e., processes for which the probability of an event per unit time interval is constant and independent of the number and timing of events which occurred in the past) or the fraction of individuals (or anything else) remaining in a system at various times after the start of an exponential decline. The mode of exponential distribution is zero and the probability of occurrence continually decreases with increased values. The skewness of an exponential distribution is two. This distribution complements the Poisson distribution which characterizes the number of occurrences per unit time and is a special case of the gamma and Weibull distributions. The exponential is less tail-heavy than the lognormal and extreme values therefore have a lower probability. It is characterized by a single parameter (β), representing the mean time between events.
gamma	The gamma distribution is widely used in environmental analysis to characterize pollutant concentrations as well as used in meteorological processes to characterize precipitation. It is also commonly used to represent the time to complete some task. The tail of the gamma distribution is not as tail-heavy (long) as the lognormal and it therefore ascribes a lower probability to extreme values than does the lognormal distribution. The gamma is typically describe by two parameters, a shape parameter and a scale parameter. When the shape parameter is 1, the distribution is equivalent to the exponential distribution.
lognormal	The lognormal distribution models quantities that are the product of a large number of other quantities (i.e., if one were to multiply a large number of random variables together, the result will tend toward a lognormal distribution). This distribution results when the logarithm of a random variable is described by a normal distribution. It is widely used in environmental analysis to represent positively valued data exhibiting positive skewness. Examples include concentrations of chemicals in environmental media and amounts of those media which are consumed, efficiencies of absorption, and rates of elimination of toxicants. The lognormal distribution has a heavier (longer) tail than the exponential, gamma or Weibull distributions. There are three common ways to parameterize a lognormal distribution: (1) arithmetic mean and standard deviation of the log-transformed variables; (2) geometric mean and standard deviation of the non-transformed variables; and (3) arithmetic mean and standard deviation of the non-transformed variables.
Weibull	The Weibull distribution is widely used in life data analysis, time to complete some task, and time to equipment failure. The Weibull distribution is less tail heavy than the lognormal and thus ascribes a lower probability to extreme events. It is typically described by two parameters, a scale parameter and a shape parameter. As with the gamma distribution, the distribution is equivalent to the exponential distribution when the shape parameter is 1,

The above information was obtained mainly from Hattis and Burmaster (1994), Vose (1996), Law and Kelton (1995), and Morgan and Henrion (1990)

[a]Note: Distributional plots, probability and cumulative density functions, interpretation of distributional parameters, formulae for important statistical terms (e.g., mean, standard deviation, etc.) are available from the literature (e.g., see Law and Kelton (1995), Vose (1996) and Evans et al. (1993))

Activity II – Estimation of Parameters

Once a candidate distribution family is selected (e.g., a lognormal distribution), we estimate the parameters of the candidate family in order to have a completely specified distribution for use in the simulation. Parameter estimation is generally accomplished using conventional statistical methods, the most popular of which include the method of maximum likelihood, probability plotting methods, and the method of moments. See Law and Kelton (1991), Evans et al. (1993), Gilbert (1987), and Vose (1996).

Parameter Estimation Methods

Maximum Likelihood Method. Probably the most often-used method for estimating the parameters of a distribution is the method of maximum likelihood. For some distribution families (e.g., normal, exponential, geometric), maximum likelihood estimators (MLEs) are well-defined values resulting from a straightforward algebraic calculation, but for others solving the equations is computationally intensive and special software is required.

There are a number of references which derive the MLE for several common distributions (e.g, Vose (1996), Ott (1995) Evans et. al. (1993)). For the purposes of this document we will simply state that the MLE for the mean and standard deviation of a normally distributed population are simply the mean and standard deviation, respectively, of the observed sample data. For the exponential distribution, the MLE for the single parameter of the exponential distribution is the mean of the observed sample data. For the geometric distribution, the MLE for the *p* parameter is $1/(\bar{x}+1)$.

Probability Plotting Methods. Probability plotting methods, sometimes called linear least square regression methods or regression on order statistics, are based on finding probability and data scales so that the theoretical cumulative distribution function plots as a straight line. The transformed data is then plotted against the linearized CDF and ordinary linear regression is performed to estimate the parameters of the fitted distribution. This method is applicable to theoretical distributions whose CDFs are expressible as a function of one or two parameters, for example, the exponential, normal, lognormal, and Weibull distributions. The following are instructions for linearizing the CDF and estimating the parameters of the fitted distribution:

> <u>For a distribution which has been hypothesized to be normal</u>
> Construct a normal probability plot with z(p) on the abscissa (the "x" axis) vs. each x_n value on the ordinate (the "y" axis)[1]. If the normal probability plot is a straight or near-straight line, this is evidence that the distribution is normal and the data are well-modeled by a normal curve. Using ordinary least-squares regression, calculate the slope of the fitted line and its intercept. The intercept is an estimate of the arithmetic mean of the distribution while the slope is an estimate of the arithmetic standard deviation of the distribution. These values should be compared with (and comparable to) the values calculated using ML method

[1] Specialized statistical software is available to create normal probability plots. Alternatively, one can create these plots using certain spreadsheet software. For example, to create a normal probability plot using Excel or Quattro Pro, first rank the observations (r_1, r_2, r_3, ... r_n) in ascending order (from lowest to highest) and assign each observation a rank (e.g, lowest observation receives a rank of 1, the next receives a rank of two, all the way to the Nth observation which receives a rank of N). For each observation, the cumulative rank is then calculated using a plotting position formula (e.g., the Weibull plotting position formula $r_1/n+1$). This can be considered similar to a percentile value except percentile values range to 100%. Next, the normal quantile is calculated for each cumulative rank: the normal quantile is the z-score associated with each percentile and can be determined using Excel's NORMSINV function. Finally, each observation's normal quantile (or z-score) is plotted on the x-axis against each observation on the y-axis.

described above.

For a distribution which has been hypothesized to be lognormal

Calculate the natural logarithms of each of the x_n values for n = 1 to N. Construct a normal probability plot with z(p) on the abscissa (the "x" axis) vs. each ln [x_n] value on the ordinate (the "y" axis) as described in the previous footnote (except than ln [x_n] is substituted for [x_n]). If the lognormal probability plot is a straight or near-straight line, this is evidence that the distribution is lognormally distributed and the data are well-modeled by a lognormal distribution. Using ordinary least-squares regression, calculate the slope of the fitted line and its intercept. The intercept is an estimate of the mean of the natural logarithms of the distribution (μ) while the slope is an estimate of the standard deviation of the logarithms (σ). These values should approximate the values for the mean and standard deviation, respectively, calculated by the following formulae:

$$\mu \quad \overline{\ln[x]} \quad \frac{\sum \ln[x_n]}{N}$$

$$\sigma \quad \sqrt{\frac{\sum (\ln[x_n] \quad \overline{\ln[x]})^2}{N \;\; 1}}$$

To calculate the arithmetic mean and standard deviations from these regression values (i.e., to define the distribution in its original terms), the following formulae are used:

$$\mu \quad e^{[\mu \;\; \frac{1}{2}\sigma^{\;2}]}$$

$$\sigma \quad e^{\mu}\sqrt{(e^{\sigma^{\;2}})(e^{\sigma^{\;2}} \;\; 1)}$$

For a distribution which has been hypothesized to be exponential

First, calculate the cumulative frequency by ranking the observations from lowest to highest as described in the previous footnote. Then, for each ranked observation subtract this quantity from 1 and take the natural logarithm of this difference. Plot this value on the y-axis vs. each individual data point on the x-axis. If the plot is reasonably straight, this is evidence that the distribution is *exponentially* distributed. Using ordinary least-squares regression, calculate the slope of the fitted line fixing the y-intercept of the regression line at the point (0, 1). The calculated slope of this line is the β parameter appearing in the exponential model [$f(x) = 1 - e^{-x/\beta}$] and should be compared with (and comparable to) the value calculated from the ML method for exponential distributions described above.

For a distribution which has been hypothesized to be Weibull

The two characteristic parameters of a Weibull distribution (i.e., the scale and shape parameters) can most easily be determined by either using dedicated statistical distribution fitting software or by plotting the data on specialized commercially-available Weibull probability paper (e.g., see Craver (1996)) . In the latter case, the Weibull scale and shape parameters can be read directly from the probability plot. For a Weibull curve (with a location parameter of 0), the scale parameter is typically represented by the 63.2 %-ile.

Weibull plots can also provide information about other potential distribution families. For example, the slope of the plotted points provide additional information about the distribution family or class with slopes of 1, 3, and 5 evidence of exponential, lognormal, and normal distributions, respectively.

For a distribution which has been hypothesized to be Beta

As with the Weibull distribution, characteristic parameters of a beta distribution can most easily be determined by either using dedicated statistical distribution fitting software or by plotting the data on specialized commercially-available beta probability paper.

For a distribution which has been hypothesized to be Gamma

As with the Weibull and beta distributions, gamma parameters can most easily be estimated by using commercially-available software or gamma probability paper.

An example of these methods using the hypothetical pesticide data is shown in Box 6.

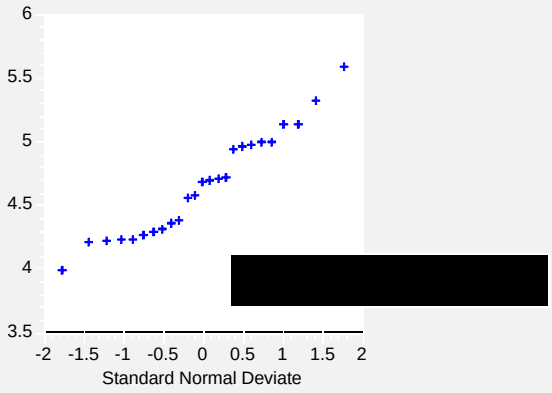
Method of Matching Moments.

The method of moments replaces each uncertain variable by its mean and variance and uses probability laws to estimate the mean and variance of the models outcome. However, the method of moments has some fairly severe limitations. For example (Vose, 1996),

- it assumes that all variables in the model are independent
- it assumes that the outcome is approximately normally distributed
- it assumes either that all variables in the model are approximately normally distributed or that the model has a very large number of uncertain variables, none of which dominates the outcome; and
- it cannot easily cope with divisions, exponents, power functions, discrete variables, etc.

BOX 6: Determination of the Appropriate Parameters for the Hypothesized Lognormal Distribution of the Pesticide Data

Having determined that a log-normal distribution is the most appropriate distribution for further analysis of the hypothetical tomato residue data, the analyst should next determine the appropriate parameters which define the distribution (i.e., the mean and standard deviation). A normal probability plot of the log-transformed values reveals a straight line with a slope of 0.4447 and an intercept of 4.65789. This intercept is an estimate of the mean of the log-transformed values (i.e., it is the ') and the slope is an estimate of the standard deviation of the log-transformed values (it is the σ')



These values are comparable to the mean and standard deviation calculated as follows:

$$\mu \quad \overline{\ln[x]} \quad \frac{\sum \ln[x_n]}{N} \quad 4.6575$$

$$\sigma \quad \sqrt{\frac{\sum (\ln[x_n] \quad \overline{\ln[x]})^2}{N \ 1}} \quad 0.4122$$

Calculating the arithmetic mean and standard deviation from the regression values in order to define the distribution in its original terms:

$$\mu \quad e^{[\mu \ \frac{1}{2}\sigma^2]} \quad 116.33$$

$$\sigma \quad e^{\mu}\sqrt{(e^{\sigma^2})(e^{\sigma^2} \ 1)} \quad 55.16$$

Thus, the most appropriate distribution to hypothesize for the hypothetical tomato pesticide residue data is a lognormal distribution with these parameters.

Activity III – Assessing Goodness of Fit

Activity III involves determining how well our selected (and now fully-defined) candidate distribution is in representing the true underlying distribution for our data. Having estimated the parameters of the candidate distributions, it is necessary to evaluate the "quality of the fit" and, if more than one candidate distribution was selected, to select the "best" distribution from among the candidates. A goodness of fit test (GoF test) is a statistical test in which the null hypothesis (H_o) is that the observed data are characteristic of a random variable with the hypothesized distribution function (e.g, exponential with a β parameter of 0.8). Unfortunately, there is no single, unambiguous measure of what constitutes best fit. Ultimately, the risk assessor must judge whether or not the fit is acceptable. This judgement should be based on a consideration of goodness-of-fit statistics as well as graphical comparisons of the fitted and empirical distributions, paying special attention to issues relevant to the analysis, e..g, fit in the lower or upper tails (but note that this is where the confidence intervals are widest). It is also important to consider the processes that generated the data and to look for probabilistic distribution models that arise from similar processes. Used in conjunction with the probability plots and statistical measures used in Activity I, GoF tests can, however, be powerful tools for verifying that a chosen distribution is at least reasonable.

Goodness-of-Fit Tests

Goodness-of-fit tests are formal statistical tests of the hypothesis that the set of sampled observations are an independent sample from the known or assumed distribution. The null hypothesis, H_o, is that the randomly sampled set of observations are independent, identically distributed random samples from a population with the hypothesized distribution. The GoF tests indicate whether the hypothesized distribution can be reasonably rejected as improbable. It is important to recognize that failure to reject H_o is not the same as accepting H_o as true These tests, taken alone, are not very powerful for small to moderate sample sizes (i.e., subtle but systematic disagreements between the data and the hypothesized distribution may not be detected); conversely, the tests can be too sensitive for large numbers of data points -- that is, for data sets with a large number of points, H_o will almost always be rejected.

Commonly used goodness-of-fit tests include the chi-square test, Kolmogorov-Smirnov test, and Anderson-Darling test. These are described further below.

Chi-Square Test. The chi-square test is based on the normalized difference between the square of the observed and expected frequencies and can be viewed as a comparison of the frequency histogram with the fitted probability density function or probability mass function. The chi-square test statistic is computed by dividing the entire range of the fitted distribution into *k* contiguous, non-overlapping intervals and counting the number of data samples falling into each interval (N_j). This count is compared to the expected number of observations in a bin. Given a sample size of *n*, the expected number of data points in the *j*th bin (*j* = 1 to *k*) is np_j where $p_j = F(x_j) - F(x_{j-1})$ .The chi-square test statistic is computed as

$$\chi^2 \quad \sum_{j\ 1}^{k} \frac{(observed \quad expected)^2}{expected} \quad \sum_{j\ 1}^{k} \frac{(N_j \quad np_j)^2}{np_j}$$

The chi-square test is highly dependent on the width and number of intervals chosen. Law and Kelton recommend selecting equi-probable bin widths such that np_j 5; D'Agostino and Stephens (1986) recommend selecting *k* equi-probable intervals where $k = 2n^{2/5}$ For example, if one had 100 data points, one might wish to form k = 13 (equiprobability) intervals. If 13 equiprobability intervals are formed for the 100 data points, then the expected number of points in each interval (i.e., the np_j) would be calculated as follows:

$$n \times \frac{1}{k} \quad 100 \times \frac{1}{13} \quad 8$$

This satisifies the criteria that each bin size be chosen such that an equal number of points (in this case, 8) numbering at least five be expected in each bin. The size of each bin width is calculated by inverting the cumulative distribution function[2]. This is best illustrated by returning to our pesticide example as shown in Box 7.

Kolmogorov-Smirnov Test. The Kolmogorov-Smirnov Test is a non-parametric test based on the maximum absolute difference between the theoretical and sample (or step-wise empirical) Cumulative Distribution Functions (CDFs). Large values of this statistic indicate a poor fit while small values indicate a good fit. Critical values for the K-S statistic depend on whether or not the parameters of the distribution are known *a priori* or have to be estimated from the data. See Law and Kelton (1992) and D'Agostino and Stephens (1986).

The Kolmogorov-Smirnov test is most sensitive around the median and less sensitive in the tails and is best at detecting shifts in the empirical CDF relative to the known CDF. It is less proficient at detecting spread but is considered to be more powerful than the chi-square test.

Anderson-Darling Test. The Anderson-Darling test is designed to test goodness-of-fit in the tails of a probability density function based on a weighted-average of the squared difference between the observed and expected cumulative densities. Additional information and critical values for Anderson-Darling statistic for the all parameters known case, and for the normal , exponential, and Weibull distributions are given by Law and Kelton (1992)and D'Agostino and Stephens (1986). Because of its relative emphasis on fit in the tails, the Anderson-Darling statistic may be particularly useful to assessors as a goodness-of-fit statistic.

[2] While these inverses can be calculated algebraically for functions with closed forms such as the exponential, use of a spreadsheet program or numerical methods may be necessary for continuous functions such as the normal, lognormal, gamma, and beta distributions. Excel© and QuatroPro© have built-in inverse functions which are called NORMSINV, LOGINV, GAMMAINV, and BETAINV, respectively, which return the value associated with any given probability. In our hypothetical pesticide example (see Box 7), the given probability is equal to 1/j for j = k down to 1, with k = 5 (i.e., 1/j = 0.2 for the first bin width, 0.4 for the second bin width, 0.6 for the third width, 0.8 for the fourth, and 1.0 for the last).

BOX 7: Equiprobability Chi-Square Test of Sample Pesticide Data

For our pesticide example, we have a total of 25 data points and desire to select *k* equi-probable intervals. We select *k* a value of 5: although the formula would yield for k a value of 7 ($k = 2(25^{2/5})=7$), we require a minimum of 5 data points per bin and thus for 25 points, 5 bins (or equiprobability intervals) are necessary. If 5 equiprobability intervals are formed for the 25 data points, then the expected number of points in each interval (i.e., the np_j) is 5 (or n x 1/k = 25 x (1/5)). With 5 bins (or intervals), the given probability is equal to 1/j for j = k down to 1 with k = 5. That is, 1/j = 0.2 for the first bin width, 0.4 for the second bin width, 0.6 for the third bin width, 0.8 for the fourth, and 1.0 for the last. The individual bin widths are calculated using Excel's LOGINV function with the assumed mean and standard deviation calculated in Activity II. The individual bin widths, observed number of points in each bin, the expected number of points in each bin, and the calculated Chi-square values are shown below:

.Calculation of Chi-Square Value for Pesticide Example Using a Lognormal (116.3, 55.2) Hypothesized Distribution

J	Interval[a]		No. Observed	No. Expected[b]	Chi-Square[c]
	Lo	Hi			
1	0	72.46	6	5	0.2
2	72.46	94.14	4	5	0.2
3	94.14	117.94	6	5	0.2
4	117.94	153.2	5	5	0
5	153.2.		4	5	0.2
TOTAL			25	25	0.8

[a] Intervals are calculated by evaluating the inverse of the hypothesized distribution at each j value. In this example, the hypothesized distribution is lognormal with an arithmetic mean of 116.3 and an arithmetic standard deviation of 55.2. Since this distribution has no closed form, the upper end of each of the 5 intervals must be evaluated with Excel (or QuatroPro) using the LOGNORMINV function with a mean (of the logs) of 4.657489 and a standard deviation (of the logs) of 0.444947 (each of which were calculated previously in Box 6).

[b] The number expected in each bin was calculated previously as n x 1/k

[c] Each chi-square value is calculated as $(\text{observed-expected})^2$ / expected. The final chi-square value is calculated as the sum of these individual chi-squared values

The degrees of freedom is given by $\nu = k - m - 1$ where k is the number of bins (or classes) and m is the number of parameters we are estimating from the data (i.e., the mean and standard deviation). From this, $\nu = 5-2-1 = 2$. The χ^2 critical value for $p = 0.1$ and 2 degrees of freedom is calculated as $\chi^2(0.9;2) = 4.6$. Since our observed χ^2 value of $0.8<4.6$, , we are unable to reject the lognormal model with an arithmetic mean of 116.3 and an arithmetic standard deviation of 55.2 on the basis of this chi-squared test of fit: the Chi-square value suggests that there is no reason to conclude that our data are poorly fitted by our hypothesized lognormal distribution.

Cautions Regarding Goodness-of-Fit Tests

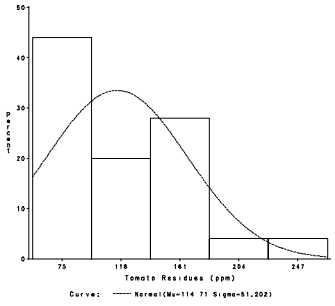
Care must be taken not to over-interpret or over-rely on the findings of goodness-of-fit tests. It is far too tempting to use the power and speed of computers to run goodness-of-fit tests against a generous list of candidate distributions, pick the distribution with the "best" goodness-of-fit statistic, and claim that the distribution that fit "best" was not rejected at some specific level of significance. This practice is statistically incorrect and should be avoided [Bratley *et al.*, 1987, page 134]. As indicated previously, Goodness-of-fit tests have notoriously low power and are generally best for rejecting poor distribution fits rather than for identifying good fits. For small to medium sample sizes, goodness-of-fit tests are not very sensitive to small (but potentially significant) differences between the observed and fitted distributions. On the other hand, for large data sets, even minute differences between the observed and fitted distributions may lead to rejection of the null hypothesis. For small to medium sample sizes, goodness-of-fit tests should best be viewed as a systematic approach to detecting gross differences.

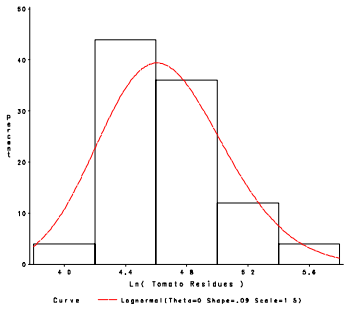
We note that there is absolutely no substitution for careful visual inspection of both the data and the theoretical distribution of the fit to the data. The human eye and brain are able to interpret and understand data anomalies far beyond the ability of any computer program or GoF tests. GOF tests may, *at best*, simply serve to confirm what the analyst has found though visual inspection. One may quite appropriately decide to retain a particular probability model despite having rejected it on the basis of GoF tests if it appears to be a good fit to the data as judged by the visual inspection of the probability plots and other comparisons.

Graphical (Heuristic) Methods for Assessing Fit

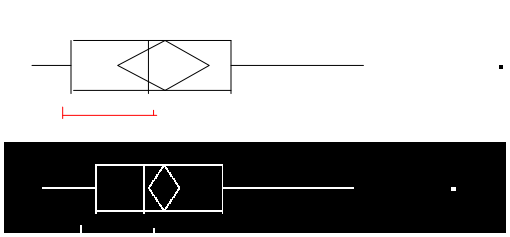
Graphical methods provide visual comparisons between the experimental data and the fitted distribution. Despite the fact that they are non-quantitative, graphical methods often can be most persuasive in supporting the selection of a particular distribution or in rejecting the fit of a distribution if one has a sufficiently large sample size. This persuasive power derives from the inherent weaknesses in numerical goodness-of-fit tests. Commonly used graphical methods for assessing goodness of fit include:

Frequency comparisons compare a histogram of the experimental data with the density function of the fitted data. Frequency comparisons must be interpreted with care since the visual comparison will depend on the number of bins used to generate the histogram of the data. Two examples of a frequency comparison are shown below for our sample pesticide data. The leftmost illustration compares the untransformed pesticide data to the normal curve while the illustration to the right compares the log-normalized pesticide residue data to the normal curve





Box plot comparisons compare a box plot of the observed data with a box plot of the fitted distribution. This is illustrated below for the sample pesticide residue data (observed) and the lognormal distribution (fitted).



Probability-Probability plots (P-P plots) compare the observed cumulative density function (i.e., the sample probability) with the fitted cumulative density function (i.e., the model probability). P-P plots are used to graphically evaluate how well the data fit a given (hypothesized) theoretical distribution, e.g. normal, lognormal, Weibull, etc. P-P plots tend to emphasize differences in the *middle* of the predicted and observed cumulative distributions, and are less sensitive than Q-Q plots to differences in the tails (where risk assessors are more frequently interested).

Theoretical Quantile-quantile plots (Q-Q plots) graph the *quantiles* of the specific fitted (or theorized) distribution against the *quantiles* of the actual data. To construct a theoretical Q-Q plot, one sorts the data in ascending order and calculates a cumulative frequency (as done for the normal probability plot) using the standard plotting formula (i.e., $r_I/N + 1$). At this point, the z value associated with this probability (or cumulative frequency) value is calculated and transformed to its original scale. In other words the quantile value associated with this cumulative probability from the *theoretical distribution* is calculated. This can be done with Excel or QuantroPro using their inverse cumulative probability functions (e.g., NORMINV, LOGINV, or GAMMAINV) or can sometimes be done analytically using an algebraic formula for distributions for which there is a closed form for the cumulative probability function (e.g., the exponential and Weibull distributions). [3] Finally, the actual data values are plotted against the values which would have been seen if the data were distributed according to the hypothesized distribution.

The theoretical Q-Q plot is used to determine how well the data set is modeled by the theorized distribution: any systematic deviations in the distribution of our sample data from the hypothesized distribution are highlighted and (ideally at least) will be readily apparent. If the graph is linear (and there are no significant systematic deviations from linearity), this is evidence in support of the data fitting the specific hypothesized distribution. Q-Q plots tend to emphasize differences in the *tails* of the fitted and observed cumulative distributions. The *deviation* of a Q-Q plot from a straight line can provide diagnostic information about the theorized distribution. For example, if the data in the upper tail fall above the quartile line and those in the lower tail fall below it, there are too few data in the tails than would be expected in the theoretical distribution (and the theorized distribution is said to be too heavy in the tails). Conversely, if the data in the upper tail fall below the quartile line and those in the lower tail fall above it, then there are more data points in the tails than would be expected in the theorized distribution (and the theorized distribution is said to be too light in the tails). Patterns in deviations from linearity can be investigated by use of a residuals plot to detect systematic departures.

[3] The theoretical Q-Q plot for the normal (and log-transformed lognormal) distributions are essentially equivalent (except for scaling) to the normal probability plot discussed earlier and constructing Q-Q plots for the normal and lognormal distributions would therefore be of little additional value.

Section III Non-Parametric Distribution Functions

Many times in Monte Carlo analyses, a non-parametric function (or empirical distribution function (EDF)) is used to characterize a model variable. In these situations, the risk assessor has determined that the data itself provides the best representation of the exposure variable. Simply put, the risk assessor has chosen to directly use the sample values to define the distribution of the exposure variable rather than represent it by a theoretical distribution fit to the data.

D'Agostino and Stephens (p.8-9,1986) discuss the advantages of EDFs. Some of the benefits of likely interest to risk assessors include:

1. *EDFs provide complete representation of the data without any loss of information.*
2. *EDFs do not depend on any assumptions associated with parametric models.*
3. *For large samples, EDFs converge to the true distribution for all values of x.*
4. *EDFs provide direct information on the shape of the underlying distribution, e.g., skewness and bimodality; EDFs supply robust information on location and dispersion.*
5. *An EDF can be an effective indicator of peculiarities (e.g., outliers)*
6. *An EDF does not involve grouping difficulties and loss of information associated with the use of histograms*
7. *Confidence intervals are easily calculated.*
8. *EDFs can be effectively used for censored samples.*

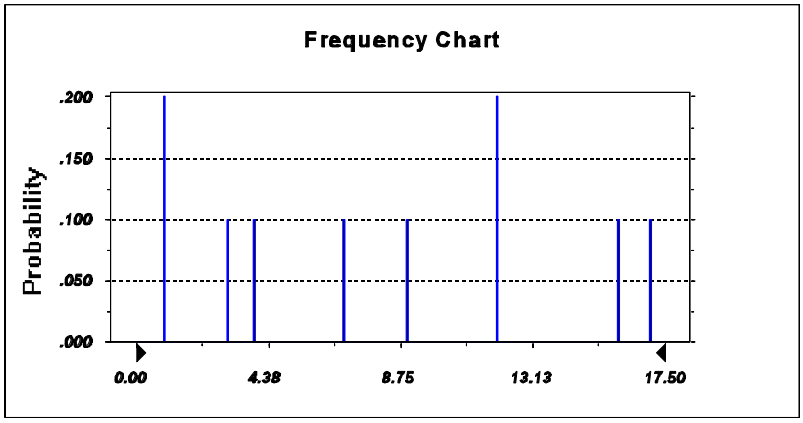
D'Agostino and Stephens also point out one of the potentially serous drawbacks to EDFs: *EDFs can be sensitive to random occurrences in the data and sole reliance on them can lead to spurious conclusions. This can be especially true if the sample size small.* In addition, we note that empirical distributions (as traditionally used) do not permit data to be generated which are *outside* the range of historically observed data and EDFs therefore tend to underestimate the probability of an extreme event.

The choice of whether or not to use an EDF in an assessment employing Monte Carlo methods is ultimately up to the risk assessor and his/her level of comfort and confidence with the data and the method. It must be remembered that EDFs (when used in the usual manner) rely solely on past observations and therefore preclude generation of data outside the historically-observed range. Monte-Carlo results generated from an EDF may produce tails that are too short and can therefore underestimate the probability of extreme events.

Below, we discuss how an EDF is defined and present several approaches used to implement EDFs.

Discrete Representation of EDFs

Given a random sample of n observations, $X_1, X_2, \cdots, X_n$,, a discrete representation of this EDF would be represented as X = $\{X_1, X_2, \cdots, X_n,\}$. These values could be used themselves directly in the simulation in what is termed a "trace-driven" simulation. In this technique, values from the raw input data are repeatedly selected in a random manner and used to calculated model outputs. For example, given the data set X = {1,1,3,4,7,9,12,12,16,17}, a discrete representation of this data set is illustrated below:



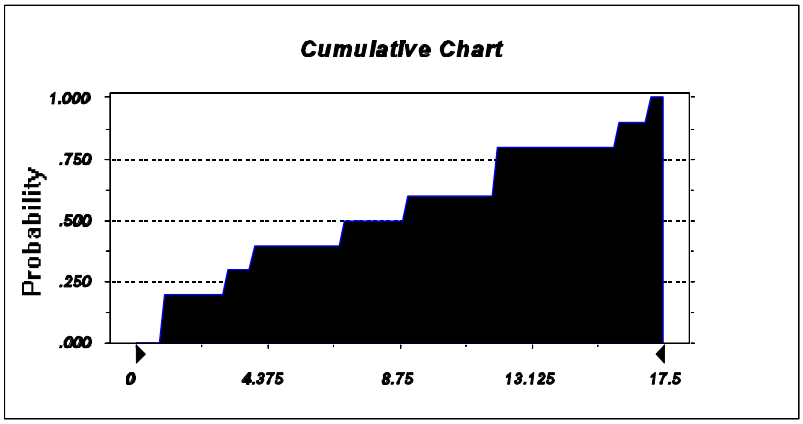
We note that with this representation no intermediate values (e.g., 2, 5, 6, 8, etc) can be generated and the simulation is limited to only those values which have historically been observed and are present in the data input set.

Continuous Representation of EDFs

Given a random sample of *n* observations, $X_1, X_2, \cdots, X_n$, sorted from smallest to largest, from a true but unknown distribution, an *empirical distribution function*, EDF, expressed on a cumulative basis may be defined as

$$prob\,(X \le x\ ) \qquad F(x\ ) \qquad \frac{number\ of\ x\ s\ \le\ x}{N}$$

For example, given the same data set X = {1,1,3,4,7,9,12,12,16,17}, the probability that X $\le$ 11 is given by F(11) = 6/10 = 0.60 since there are 6 samples with values less than or equal to 11 and there are ten samples in the entire data set. This formulation of the EDF presents some problems since all values of x^* in the range $9 < x^* \le 11$ have the same probability (called *constant interpolation*), i.e., prob ($X \le 10$) = 6/10, prob ($X \le 10.5$) = 6/10, prob ($X \le 11.5$) = 6/10, and so on. Defined this way, the EDF is a step function with abrupt jumps at the sample values as illustrated below:



The EDF is then expressed as

$$F_n(x) \quad \begin{cases} 0 & x < x[1] \\ \dfrac{k}{n} & x[k\ \ 1] < x \le x[k] \qquad for\ k \quad 1,2, \quad ,n \\ 1 & x > x[n] \end{cases}$$

where x[0] is set to zero. As with the discrete representation, values below the sample minimum and beyond the sample maximum cannot be generated. However, unlike the discrete representation, <u>any</u> value between the maximum and minimum can be generated.

Linear Interpolation of Continuous EDFs. It may be unsettling to define the EDF as a step function with abrupt jumps at certain values and so interpolation is often used to estimate the probabilities of values in between sample values. Generally , for values between observations, i.e., $X_{k-1} \le x < X_k$, linear interpolation is used, although higher order interpolation is sometimes used. The EDF for linear interpolation between sample values is simply

$$F_n(x) \quad \begin{cases} 0 & x < x[1] \\ \dfrac{k}{n} \quad \dfrac{x - x[k]}{n(x[k] - x[k-1])} & x[k-1] < x \le x[k] \qquad for \quad k \ \ 1,2, \quad n \\ 1 & x > x[n] \end{cases}$$

Extended EDF. The linearly interpolated EDF cannot produce values beyond the values in the data sample. This may be an unreasonable restriction in many cases. For example, the probability that a previously observed largest value in a sample based on *n* observations will be exceeded in a sample of *N* future observations may be estimated using the relationship $prob = 1 - n/(N + n)$. If the next sample size is the same as the original sample size, there is a 50% likelihood that the new sample will have a largest value greater than the original sample's largest value. Restricting the EDF to the smallest and largest sample values may produce distributional tails that are too short.

In order to get around this problem, one may extend the EDF to include plausible minimum and maximum values. The extended EDF expands the linearly interpolated EDF by including a user-defined absolute minimum, x_{min}, and absolute maximum, x_{max}, which are beyond the data sample.

$$F_n(x) \qquad \begin{array}{ll} 0 & x < x[0] \\ \dfrac{k}{n\ \ 1} \quad \dfrac{x - x[k]}{(n\ \ 1)\,(x[k] - x[k-1])} & x[k-1] < x \le x[k] \quad for \quad k\ \ 1,2, \quad n\ \ 1 \\ 1 & x > x[n\ \ 1] \end{array}$$

where $x[0] = x_{min}$ and $x[n+1] = x_{max}$.

References and Suggested Readings

American Industrial Health Council, *Exposure Factors Sourcebook*, May, 1984.

P. Bratley, B. L. Fox, L. E. Schrage, *A Guide to Simulation*, Springer-Verlag, New York (1987).

D.E. Burmaster and E.A.C. Crouch, "Lognormal Distributions for Body Weight as a Function of Age for Males and Females in the United States 1976-1980. Risk Analysis, **1**, 499-505 (1997)

J.S. Craver, *Graph Paper From Your Computer or Copier*, Fisher Books. 3rd. Ed., (1996)

R.B. D'Agostino and M.B. Stevens, Goodness of Fit Techniques, Marcel Deker (1986)

Decisioneering, Inc. (1996) Crystal Ball Version 4.0 User Manual.

M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, John Wiley & Sons, New York (1993).

R. O. Gilbert, *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York (1987).

L. C. Hamilton, *Regression with Graphics - A Second Course in Applied Statistics*, Duxberry Press, Belmont, CA (1992).

D. Hattis and D.E. Burmaster, "Assessment of Variability and Uncertainty Distributions for Practical Risk Analysis", *Risk Analysis* **14**, 713-730 (1994),

D. R. Helsel and R. M. Hirsh, *Statistical Methods in Water Resources*, Elsevier, New York (1992).

A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, Chapter 6, 325-419 (especially 356-404), McGraw-Hill, New York (1991).

J. Lipton, W. D. Shaw, J. Holmes, and A. Patterson, "Short Communication: Selecting Input Distributions for Use in Monte Carlo Simulations", Regulatory *Toxicology and Pharmacology*, **21**, 192-198 (1995).

M.G. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press (1990)

W. Nelson, *Applied Life Data Analysis*, John Wiley & Sons, New York (1982).

Wayne R. Ott, *Environmental Statistics and Data Analysis*, Lewis Publishers (1995).

Palisades. *@RISK Users Manual*

A.M. Roseberry and D.E. Burmaster, "Lognormal Distributions for Water Intake by Children and Adults," *Risk Analysis* **12**, 99-104 (1992).

S. H. C. du Toit, A. G. W. Steyn, R.H. Stumpf, *Graphical Exploratory Data Analysis*, Springer-Verlag, New York (1986).

M. B. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data", *Biometrika*, **55**(1), 1-17, (1968).

U.S. EPA, *Guidance for Data Quality Assessment: Practical Methods of Data Analysis* EPA QA/G-9, EPA/600/R-96/084, July, 1996. Available on-line at http://Earth2.epa.gov/ncerqa/qa/docs/epaqag9.PDF

U.S. EPA *Exposure Factors Handbook* August, 1996. DRAFT. EPA/600/P-95-002a,b,c

D. Vose, *Quantitative Risk Assessment: A Guide to Monte-Carlo Simulation Modeling*. John Wiley and Sons (1996)

ATTACHMENT 4

ATTACHMENT 4: Memorandum entitled "Final Office Policy for Performing Acute Dietary Exposure Assessment", D. Edwards, June 13, 1996

June 13, 1996

MEMORANDUM

SUBJECT: Final Office Policy for Performing Acute Dietary Exposure Assessment.

FROM: Debra Edwards, Ph.D., Acting Deputy Director
Health Effects Division (7509C)
Office of Pesticide Programs

TO: CBTS, CBRS, DRES, and RCAB Staff

Enclosed is a copy of the final policy which describes how OPP currently performs acute dietary risk assessment (Tiers I and II). The policy also outlines proposed future refinements to our current policy (Tiers III and IV). The policy has been reviewed by the SAP in September, 1995.

Attachment

cc: Stephanie Irene
Penelope Fenner-Crisp
Lois Rossi
Steve Johnson
Anne Lindsay
Bill Jordan
Richard Schmitt
HED Branch Chiefs

ACUTE DIETARY EXPOSURE ASSESSMENT
Office Policy
June 1996

Purpose

The purpose of this guidance document is to outline how OPP performs acute dietary exposure assessments. This document is for internal use and reflects current policy, reviewed by the SAP in September 1995.

Background

OPP intends to use a tiered approach to determine acute dietary exposure associated with pesticide use. The steps in the analysis proceed from more to less conservative assumptions. The tiered approach is considered the most efficient means of exposure assessment both for the Agency and Industry, matching the level of Agency and Industry resources used to the level of risk concern. For Tiers 1 and 2, no additional data will be required of the registrant; the registrant will be required to mitigate any unacceptable risk from Tier 3 analyses and, at their own option, may generate additional single serving size[4] (Tier 4) residue data. Analysis proceeds only to the step at which no risk concern is indicated. Selection of an MOE that triggers concern is tied to the nature of the adverse effect under consideration. Both individual and population risk will be considered in regulatory decision-making.

Currently, the Agency performs only Tier 1 and 2 acute exposure assessments. The future addition of Monte Carlo[5] analysis capability will allow the use of Tier 3 and 4 acute exposure assessments.

---

[4] "single serving size" - individual pieces of a commodity for which one discrete piece constitutes a serving (such as a single apple or single banana)

[5] A Monte Carlo analysis creates a joint distribution of two variables, in the case of DRES, by randomly pairing a distribution of residue chemistry data with a distribution of food consumption data, to create a representation of the actual exposure distribution.

Tiered Acute Exposure Assessment Overview

Tier 1, using a single high end residue estimate and a distribution of consumption data, is inexpensive and the least resource intensive but gives only an upper bound (worst-case) estimate of acute exposure.

Tier 2, the same as Tier 1, except using a single average residue data point for commodities which are typically mixed, requires minimally more effort than Tier 1, but provides a more realistic estimation of exposure by considering average anticipated residues for food forms that are typically mixed prior to consumption.

Tier 3, using a distribution of residue data points as well as a distribution of consumption data points, requires additional Agency review time, but provides a more realistic estimation of acute exposure than Tier 2.

Tier 4, using a distribution of residue data points from single serving size samples, is the only method which requires additional residue data from the registrant. It requires additional and expensive residue data, extensive Residue Chemistry and DRES review time, but provides the most representative exposure picture. However, it may not provide a lower exposure estimate than Tier 3.

Procedure

Tier 1 uses a single high end residue estimate (usually the tolerance) together with a distribution of consumption values to estimate single-day exposure. This tier assumes the following:

- All commodities which have a tolerance for a pesticide contain tolerance level pesticide residues (or the highest residue found in a field trial).

- If residue data for the edible portion are reported in the field trials, the residue estimate for the edible portion is taken from the highest residue found in field trials conducted at the maximum use pattern on the label.

- The tolerance, or maximum legal level of a pesticide in or on a human food or animal feed commodity, is derived from the field trial composite sample[6] exhibiting the highest residue.

- 100% of the crop is assumed to be treated.

- Tolerances/residue estimates for "all raw agricultural commodities" in food handling establishments will be excluded from the analysis.

---

[6] Composite samples are numerous pieces of a commodity which are blended together prior to analysis (such as 12 large potatoes or 24 lemons). Composite samples are collected in field trials because FDA monitoring for tolerance enforcement is done using composite samples, and the primary purpose of a tolerance is as an enforcement tool. Guidelines on the minimum sample sizes are outlined in the Codex "Guidelines on Minimum Sample Sizes for Agricultural Commodities from Supervised Field Trials for Residue Analysis", ALINORM 87/24A (1987).

- Currently the exposure to consumers[7] only is calculated; non-consumers are excluded from the analysis.

Tier 2 is the same as Method 1 for commodities which are commonly consumed as a "single serving size", or cannot be assumed to be mixed during processing; e.g., apples, oranges, pears, bananas, potatoes. For food forms that are typically mixed prior to consumption [grains (e.g. rice) and grain products, oils, sugars, most juices, tomato products (paste, puree, and juice), dried potatoes, soybeans, peanuts, mint oils, milk, wine, and sherry], an average anticipated residue from field trial data or 95th percentile residue from monitoring data is combined as above with a distribution of consumption data to estimate exposure.

- The high end residue for commodities consumed as a "single serving size" is determined the same way as it was in Tier 1.

- The residue estimate for raw agricultural commodity food forms that are typically mixed (e.g., rice, dry beans) is determined by averaging the residue data from field trials conducted at the maximum use pattern on the label. Alternatively, the 95th percentile residue from monitoring data may be used.

- The residue estimate for food forms (processed foods) that are typically mixed is determined by using the average residue found in field trials conducted at the maximum use pattern multiplied by the average processing factor determined in processing studies. Alternatively, the 95th percentile residue from monitoring data may be used. For processed food forms that may be derived from a limited geographic region (individual farm, county), the highest average field trial (HAFT) should be used.

- In calculating the average residue, if the residue level of the pesticide falls below the estimated limit of detection (LOD) of the method, the limit of detection will be assigned. If the residue level of the pesticide falls between the estimated limit of detection of the method and the limit of quantitation (LOQ) of the method (point at which quantitative results may be obtained with a specified degree of confidence), the residue will be estimated to be the LOQ.

Tier 3 combines the entire distribution of residues from field trials (composite samples) with the entire distribution of consumption data to estimate a distribution of exposure (convolution of distributions using the Monte Carlo method). Tier 3 allows the following:

- A distribution of residue data points is included for all possible commodities, which is more realistic than a single point estimate.

- If residue data for the edible portion are reported in the field trials, the distribution of residues from field trials conducted at the maximum use pattern on the label is used. If residue data on the edible portion are <u>not</u> available, the residue data points for the raw commodity may all be multiplied by the

[7] An individual on a given day is defined as a "consumer" if he consumed one or more of the foods for which a prior tolerance exists for a pesticide. For example, if only 1000 out of the 90,000 consumption data records include consumption of strawberries, then only those records would be used in the analysis. This was originally done to be protective of people actually consuming the commodities of concern.

average processing factor to determine the distribution of residues for the edible portion of the raw commodity.

- For commodities which are typically mixed, in general, a point estimate will be used i.e. the average residue from field trials, multiplied by the average processing factor or the 95th percentile residue from monitoring data. As in Tier 2, the HAFT should be used for processed food forms that can be derived from a limited geographic region. However, distributions of residues will be used when it is necessary to further refine the analysis.

- Percent crop treated data are included in the equation, by assigning a probability that the residue level could be zero. However, consideration must be given to the possibility of regional outliers, where a higher percent of the crop is treated and marketed primarily within that region. Imported crops are assumed to be 100% treated unless better data are available.

- The total population is included in the assessment (consumers and non-consumers). This allows population exposure comparisons between analyses for different commodities and different pesticides.

- As data become available on variability inherent in composite sampling, monitoring data may be used in acute dietary exposure analysis for all foods.

Tier 4 (optional) combines the entire distribution of residues from a well designed, statistically valid market-basket survey (single serving-size samples, i.e., not composited), with the entire distribution of consumptions to estimate the distribution of risks (also Monte Carlo method).

- The entire distribution of residue data points from the specially conducted market basket survey is used as residue estimates for commodities consumed as single servings. Single serving size samples are collected in these special surveys. The edible portion of the commodity is analyzed. Alternatively, the estimates for the whole commodity may be modified by multiplying by processing factors to determine the edible portion of the commodity.

- Individual serving size samples, e.g., individual apples, and a corresponding composite sample collected from the identical sampling site both should be analyzed. The concurrent analyses of individual units/serving portions and the corresponding composite may provide a basis on which later, independent monitoring data using composite sampling may be used by the Agency to assess acute dietary exposure. Additional guidance on the conduct of special surveys for Tier 4 will be provided as needed for each study.

Procedures for DRES portion of Acute Exposure Analysis

- Current DRES acute analysis uses the individual person-day data from a survey of food consumption, in which approximately 30,000 people were surveyed for 3 days each, approximately 90,000 person-day records. Each record contains information on the consumption of the 376 "raw agricultural commodities" (RACs) which make up the standard list of RACs in the DRES system. Data are recorded as grams of RAC eaten per kilogram of bodyweight on that day. If a person did not consume a particular RAC, the RAC is given a "missing value" in the database.

- Acute analyses use the person-day consumption data in the fundamental formula: Exposure = Residue x Consumption

$$E_i = R_i \times C_i \times 0.001$$

where

E = exposure from the pesticide on RAC 'I', in milligrams pesticide/kilogram bodyweight/day,

R = the chemical residue on RAC 'I', in mg pesticide/kg RAC,

and

C = the consumption of RAC 'I', in g RAC/kg bodyweight/day.

0.001 = conversion from grams food to kilograms food

- In acute analyses, exposure is summed across all RACs for each person, and the distribution of exposures across the population is plotted on a histogram or table. The statistical weights assigned to each individual in the survey are taken into account.

- Acute exposure is expressed as a margin of exposure (MOE). The MOE is calculated using the equation,

$$\text{MOE} = \frac{\text{NOEL}}{\text{exposure}} = \frac{\text{NOEL}}{\text{residue x consumption}}$$

- The magnitude of risk is generally estimated by comparing the exposure value to the highest dose level known *not* to cause effects (NOEL or other appropriate endpoint). Subgroups are of concern in acute analyses.

- Selection of an MOE that triggers a risk concern should be tied to the nature of the adverse effect under consideration and the type of study from which the NOEL is taken. Effects that are reversible may be regulated less stringently than those which are irreversible and life threatening. Dose-Response information is also a consideration.

- The acute DRES analysis does not take into account food handling establishments. We believe that the underlying assumption, that all commodities that are consumed on any given day will contain tolerance level residues of pesticides from a food handling establishment, is unrealistic. Residues resulting from pesticide use in food handling establishments are not likely to result in incidental contamination of all foods at tolerance levels on a uniform and consistent basis and not all foods consumed by an individual in a day are likely to have come from a food handling establishment.

- Future acute DRES analyses will reflect the total population and not consumers only. The population of "consumers" will be different for every analysis and the comparison of different chemicals is not appropriate with the present procedure. The inclusion of the total population in future analyses will permit comparison between commodities for any given pesticide, and also permit analysis of alternatives by comparing the risk picture between pesticides.

- Percent crop treated data will be included in Monte Carlo analyses in the form of 'zero' residue data points in the appropriate proportion, when field trial data are used. This will result in exposure estimates more nearly reflecting the actual exposure.

- Residues in water should be included in the DRES analysis, by including the MCL or monitoring data, as available. However, regional variations in exposure must be considered when characterizing population risks.